AGI Consciousness

Piotr Bołtuć

University of Illinois, Springfield. USA
Warsaw School of Economics, Poland
pboltu@sgh.waw.pl

Abstract
AI can think, though we need to clarify definition of thinking. It is cognitive, though we need more clarity on cognition. Definitions of consciousness are so diversified that it is not clear whether present-level AI can be conscious – but this is primarily for definitional reasons. To fix this would require four clusters: functional consciousness, access consciousness, phenomenal consciousness, hard consciousness. Interestingly phenomenal consciousness may be understood as first-person functional consciousness, as well as non-reductive phenomenal consciousness the way Ned Block intended. The latter assumes non-reducible *experiences* or *qualia,* which is how Dave Chalmers defines the subject matter of the hard problem. I pose that the Hard Problem should not be seen as the problem of phenomenal experiences, since those are just objects in the world (specifically, in our mind). What is special in non-reductive consciousness is not its (phenomenal) content, but its epistemic basis (the carrier-wave of phenomenal qualia) often called *the locus of consciousness.* It should be understood through 'subject that is not an object'. This requires a complementary ontology of subject and object, developed by [Russell 1921], [Nagel 1979] and presented at [Boltuc 2019]. Reductionism is justified in the context of objects, including the experiences (phenomena), but not in the realm of pure subjectivity – such subjectivity is relevant for epistemic co-constitution of reality [Fichte, Husserl, *maybe* Kant *for whom the subject was active, so it was a mechanism and mechanism are all objects*]. Pure epistemicity is hard to grasp; it transpires in second-person relationships with other conscious beings [Buber, Levinas] or *monads* [Leibniz, Conway]. If AGI is to dwell in the world of meaningful existences, not just their shadows, as the case of Church-Turing Lovers highlights [Boltuc 2017], it requires full epistemic subjectivity, meeting the standards of the Engineering Thesis in Machine Consciousness [Boltuc 2009, 2012, 2007].

Keywords: machine consciousness; AI consciousness, AI thinking, AI cognition, AGI-consciousness; Artificial General Intelligence; phenomenal consciousness; p-consciousness, functional consciousness, f-consciousness, a-consciousness, h-consciousness; hard consciousness; subject-object: pure subject; Thomas Nagel; *View from Nowhere*: The Engineering Argument in Machine Consciousness; Church-Turing Lovers

1 AI Thinking

Turing was right that, eventually, there would be no cognitive activity that humans conduct better than computers. Lady Lovelace, to whom Turing refers as a potential critic, had a good point that "The Analytical Engine has no pretensions to originate anything. It can do whatever we know how to order it to perform" [Turing 1950]. Nowadays, it is not the case that all AI engines are 'analytical engines' in that sense. We also have weaker and weaker reasons to view human thinking as a process dominated by predicative logic. For instance, genetic algorithms do create new useful information through the cognitive architecture consisting of a randomizer focused on limited transformations of the image of

preexisting reality and an AI 'critic' that selects those permutations of reality (designs) that satisfy useful functional specifications [Thaler 2014] There are various designs relying on highly complex stochastic process occurring *at the edge of chaos* [Goertzel 2006]. Just like AI, humans seem to think and create primarily at the unconscious level [Libet], while language is an evolutionary device helpful in social transmission of science, and less so in individual creativity, including mathematical and scientific discovery, where intuitively grasped paradigms and complex dynamic visualizations play a large role [Boltuc 2019b, 2019c, Thaler 2019].

Some skeptics still search for cognitive domains *in principle* inaccessible to AI. If one understands that all functions, including the cognitive ones, are guided by algorithms, which, *qua* algorithms, are necessarily replicable in AI (physical interpretation of the Church-Turing Thesis [Deutsch]), it simplifies the debate immensely. There is no good reason to expect any cognitive functionalities unique to animal, in particular human, brains (because they would have to be non-algorithmic, and I am not even sure what that would mean).

One may think of motivations as separate from cognitive functions, but this is flawed. Copious work on robot emotions and motivational structures, as simple as structured *rewards*, demonstrates such distinction to be irrelevant. Hobbes' explanation of the functioning of mind, including his analogy between magnetic forces and *attraction* or *repulsion* in animals, explained such mistakes out even before the Babbage computer.

The triumph of those reductive explanation applies to all human cognitive activities (thinking, broadly understood); thus, there is strong propensity to extrapolate that reductive explanations also apply to the area of consciousness. It does not help that the term consciousness is ill-defined, especially at the level of explanation familiar to the general audience (which in this case means nearly everybody, except for a handful of philosophers of mind). And even the 'experts' tend to disagree, and in vague terms at that. Hence, my first task is to help define consciousness for AI.

2 Consciousness for AI

What is the difference between consciousness, cognition and thinking?

Proposing regulative definitions: It is helpful to view **thinking** information processing that increases the likelihood of getting from true premises to true conclusions (mistaken thinking is a procedure likely to attain this goal that fails for some formal or contingent reason). Hence, not all thinking has to be conscious; in particular, AI engines do not have to be conscious in order to perform their functions (which often rise to the level of *thinking* in the above sense.)

Cognition can be defined as processing of sensory inputs, assimilated in cognitive schemes. Cognitive architectures, biological or artificial, that do such processing are cognitive engines.

Consciousness is more elusive since the notion covers several clusters of ideas. In general, it is a high cognitive function, which may involve visualization or similar processes pertaining to different sensory-based imagination [Boltuc 2007], but this is not quite a definition, merely a pointer. Analysis of different cognitive clusters within the idea of consciousness turns out more productive.

The types of consciousness I find focal are F-, P-, A- and H-consciousness. **Functional-consciousness** is whatever performs conscious functions; or, whatever behaves as if it had consciousness. This definition's theoretical weakness is that it refers to the sort of pointer definition I used above. But practically this problem dissipates since we all can single out the set of advanced cognitive functions that are f-conscious. *People disagree what some of those functions are, but we are not to involve such level of detail.*

Phenomenal-consciousness is "experience; the phenomenally conscious aspect of a state is what it is like to be in that state" [Block 1995]. In terms of denotation this definition is identical with Chalmers' definition of consciousness in the *Hard Problem of Consciousness* [Chalmers]. The difference lies in

connotations since for Chalmers there is a methodological gap between scientific methods and possible explanation of phenomenal experience (following *Leibniz mill*), whereas for Block such methodological gap may be illusionary, resulting from limitations of past, or also current, science.

Block's attempts to situate phenomenal consciousness as explainable by the sciences has a kick to it. His definition of phenomenal consciousness can be read in strictly functional terms, so that it pertains to cognitive functions of a robot [Franklin *et. al*], which is correct *de dicto*, but does not seem consistent with Block's intended connotation.

Access-consciousness occurs "when you have an episode of phenomenal consciousness and it is available to your cognitive systems" [Block 1995]. One may say, in terms of Baars' global workspace, that it accesses content that is 'globally' broadcast in the system. [Nathaniel]. It is important for Block to emphasize that one may have phenomenal consciousness that is not access consciousness [Block 1995, Block *recent*] since some phenomenal experiences may not be attended to. On the other hand, access consciousness is always phenomenal since there is no other content one could attend to.

Hard-consciousness is the sort of consciousness defined in the Hard Problem, and actually in the Reformed Hard Problem. Hard-consciousness is my concept, based directly on Chalmers' notion of the hard problem [Boltuc 2007]. For Chalmers the hard problem of consciousness is the problem of experience and the hard consciousness is the kind of consciousness that has those experiences. Isn't it just Block's phenomenal consciousness? Well, denotation may be the same – denoting the stram of phenomenal qualia. Yet, for Block that consciousness seems to have more epistemic machinery attached to it – ways we interact with reality, we can investigate it etc., which is why Franklin's reading of it as just first-person functional consciousness is technically correct, though unintended by Block. But such reading of Chalmers' notion of experience resulting in the hard problem of consciousness would disallow Franklin's reading – the epistemic machinery attached to this notion by Block may not change its denotation but it does extend its connotation into some first-person functional domain. Chalmers' notion of the hard problem as the problem of experience, epistemically more detached, is conceptually reduced; thus it is a different notion, however subtle the difference may be.

This is all ancient history to me since I think we ought to reformulate the Hard Problem. The abovementioned definition of hard-consciousness includes the notion of the Reformed Hard Problem of Consciousness[*], to which we now proceed.

my task is to help define consciousness for AI.

3 The Hard Problem beyond Phenomena

What is the gist of the Hard Problem of Consciousness?

The red herring of non-reductive consciousness are secondary qualities, qualia, phenomena. But think of a phenomenal object, an object in your thoughts, or the object of your perception viewed within your stream of consciousness is an **object** nevertheless. Phenomenal objects can be *objecivised*. Just like in the Clark-Chalmers extended mind hypothesis where your notebook may belong to your mind extended far beyond one's skull, the opposite is true as well. Objects internal to your mind, those *in the skull*, may be externalized in a straightforward way. Take the experiments by Jack Gallant's group of putting images that you see (and recently they whole phenomenal movie you see in your mind) directly from one's visual cortex and put it on a computer screen. Your visual cortex is a bit like a biological camera and you can get the recordings directly from it. There is also developing research on reading thoughts and other non-visual information from various areas of one's brain; good bye the insurmountable tower of privileged access. Our brain is an object and so are its products, including phenomenal images in our mind. Nothing is epistemically private anymore.

[*] This has little to do with Block's 'harder problem of consciousness'.

Does this lead us towards reductive materialism? Very much so, since those little bastions of *epistemic privacy* and kind of *non-physical content* needed to be gone a long time ago and they are getting to be gone now. This leads towards reductive materialism as far as the objects go; reductive it is, but not all the way. What's left is the *locus of consciousness* [Shalom], the very stream of one's awareness in which the content (always built only of objects, phenomenal or otherwise) becomes transparent to the mind. The way to avoid being overly poetic, and vague, about it is to go back to early Nagel, and his view on 'the subjective', but there is a reason why he seems to have abandoned this early, promising project: This account is very hard to make within the post-Locke, post-Hume Anglo-American tradition. It is easier to formulate in the tradition of classical German philosophy, from Leibniz, thorough Kant and Fichte all the way even to Husserl, but this philosophical tgradition has been relegated to the studies of the history of philosophy and, as a *living philosophy*, it is misunderstood and barely alive.

In his first major book [Nagel 1979] the author reprinted his now very broadly read article "What is it like to be a bat?" [Nagel 1974], which gave rise both to Block's view on phenomenal consciousness and Chalmers' definition of the Hard Problem. Imagine that you want to imagine what it is like to have the other senses, such as echolocation, that we are barely acquainted with. It is impossible to do. In the book this paper (Chapter 12) is just an intuition pump aimed to clear the way for the crucial chapter 14 "Subjective and Objective", which is new to the book. The latter chapter is also the gist of *The View from Nowhere* [Nagel 1986], the book viewed for many years as his masterpiece. Later, due largely to the highly skeptical critiques by McGinn, Nagel's main complex argument went out of fashion whereas its propedeutical version – the 'what it is like' paper – took the center stage.

3.1 Complementary Philosophy

Nagel's early philosophy was permeated by the seemingly paradoxical question: *How is it possible that I am one of the beings, among all the beings in the objective world?* It does not matter whether there are identical beings, whether I share continuity and connectedness (physical or mental) with any beings, since even if they and me are very much alike, there is still an open question, which one of them is me. Being me is the only way to experience oneselve's phenomenal experiences, even if those could be shared by some objective means with the others. According to Nagel, the problems such as *personal identity, mind and body, free will or agent-centered morality cannot be detached* 'from the subjective point of view on which they depend for their existence' [Nagel 1979 p. 213].

While in his early work[†] at some points Nagel sounds a bit like a dualist – e.g. when asking "how one can include in the objective world a **mental substance**[‡] having subjective properties" -- his main focus is on keeping room for the two complementary perspectives, subjective and objective "because the same individual is the occupant of both viewpoints' [op. cit. p. 208]. We live in the two mutually irreducible perspectives, subjective and objective. Both, 'idealism and its objectifying opposite' share 'a conviction that a single world cannot contain both, irreducible points of view and irreducible objective reality' [op.cit. p. 212]. Actually, the early version of Russellian monism in his *analysis of Mind* [Russell 1921] presents this exact same complementarity of the two perspectives, objective and subjective. But a few years later, in his *Analysis of Matter* [Russell 1927] Russell's view becomes a version of anomalous materialism, and so does Nagel's, overall, in his main book [Nagel 1983]. We may see the object and subject as two different viewpoints on reality, that play a complementary role [Boltuc 2009, sec. 4].

Interestingly, Nagel is good at avoiding the claim of uniqueness of (or exclusive access to) the content of one's first-person consciousness[§]: "What is more subjective is not necessarily more private", he explains since subjective experiences are, in some sense public property [Nagel 1970 p.207]. He also

[†] And also in his late and much less philosophically appealing book [Nagel 2012].
[‡] My **bold** font.
[§] Unlike most of his followers.

makes the following remark: "subjective aspects of the mental can be apprehended only from the point of view of the creature itself (perhaps taken up by someone else)" [op. cit.p.201]. The puzzling point is the parenthetical remark. Did Nagel make a simple point, made multiple times in [Nagel 1983], that we can take somebody else's point of view, just like several pictures can be taken from the exact same location? But this would be his argument for objective aspects of the subjective view. This does not seem to pertain to the 'subjective aspects of the mental'. Perhaps this could be seen as anticipation of something like Chalmers' *dancing qualia argument*: the idea that, provided the right level of neuroscience, we may be able shift observers from the insight of one brains to the other [Chalmers j]. For applications to verifiability of the first person experience see [Boltuc 2010, Schneider 2017]. This last issue is important since non-private settings of the first-person consciousness seem like the future of studies on first-person consciousness, where Galant's experiment is just the beginning. This is because scaffolding of one's brain is as objective as any other objects in the universe; the only puzzling thing is what aspect or part of the brain gives us the first-person feel (and conscious understanding) of being there.

3.2 Subject not an Object

The complementary view on subject and object implies that the subject is not an object of any kind, not even a substance[**] -- this is like putting all the objects to one side of the equation to help us see more clearly what is left. I have discussed this issue in some detail lately [Boltuc 2009a] so here let me just mention a few points. "The most reduced definition of pure epistemic subject leaves such subject with no direct predicative features (…) we can predicate *about* it, the way we do in the present sentence, but we cannot predicate *of* it, in the narrow sense of providing a direct description. By predicating *about* the epistemic subject we use a meta-level of reference." Such subject is only the condition of first-person (non-reducible, I presume) epistemicity – the beginning of the first-person perspective. It can be ontologically explained by accounts such as non-reductive materialism, double aspect theory, neutral monism [Russell 1021] but hardly through dualism, since the latter requires mental substance; mental substance is also an object and would violate the criterion that subject is not an object in ontological sense.

4 AGI and non-reductive subjects

How is the above philosophy of pure subject related to AGI? First, it shows how AI consciousness we have is not quite the kind of first-person consciousness that we assume alive people and animals have

Second, more interestingly, it points to what would have to happen for robots to have first-person non-reductive consciousness.

4.1 The Engineering Thesis

The Engineering Thesis in Machine Consciousness is the idea that if something characterizes humanoid brains it could, in principle, be engineered (or bioengineered) in AI. Remember that we are within the realm of science and engineering, whether accounted for by non-reductive materialism, by panpsychism where the mental is just one substance about the many material substances (in the spectrum from Ann Conway to David Chalmers), complementary view [Russell 1921, Nagel 1979] or some sort of double aspect theory [Feigl]. Of course, non-naturalistic worldviews, or those forms of naturalism quite different from contemporary post-materialistic world-view many scientists hold, would impose

[**] This is why Nagel's mention of mental substance [Nagel 1979 p. 201] was particularly unfortunate.

further ontological or theological conditions, but we abstract from those conditions, which does not amount to rejecting or being in any way partial for or against such views – those further conditions may be added to our bare-bone model as needed. Here is the bare bone conceptual engineering model:

If human beings and other animals have first person consciousness, then neuroscience or another discipline should eventually discover how it is generated. If so, we should eventually be able to build generators of first-person consciousness [Boltuc 2012, 2009, 2007]. Practical or ethical issues would be important once we reach proximity of realizing such model, but they are premature at this early conceptual phase.

5 Subsymbolic

Symbolic structures can be viewed as more advanced in AI than subsymbolic since they allow full human readability [Kelley]. But subsymbolic structures are incompatibly richer. Human readability should not be viewed as the gold standard of information processing [Sanz], though it is a standard of control in limited capability AI environments. Multifarious gestalts [Boltuc 2018] and what I call Spinozian phenomenology able to grasp non-human *multitudinal* qualities of experience [Boltuc 2019…] at the subsymbolic level convey incomparably richer information than predicative language.

The same seems true about human thinking since Libet's experiment revealed that consciousness is only the tip of the iceberg of human cognition. Symbolic thinking, especially linguistic communication, is very strongly linked with phenomenally conscious activity. Hence, Libet leads us to question dominance of linguistics in human mind – though it may still be a convenient way of transferring intellectual culture, especially history and science. Poetry uses language largely in subsymbolic manner, so do creative music and visual arts. They are largely subliminal and in formal terms subsymbolic. AI is behooved by using this level of information processing as dominant.

6 AGI's Culture

We should be proud of nearing, slowly but assuredly, origination of Artificial **General** Intelligence. Just like builders of the pyramids or those who discovered fundamental laws of geometry, dynamics or quantum mechanics would rightly be proud of extending horizons of humanity, those involved should be proud of transcending the capabilities of human mind. Fear comes from excessive prudence, which is the opposite of success, glory or progress. Fearful roaches would have never involved into humans, so to say. Posthumanist speculations aside (since this sort of futurology barely ever works in a long run), we should be primarily proud, not primarily worried.

When Mori discovered the Uncanny Valley effect, he was preoccupied with quite imperfectly humanoid robots invading human spheres of privacy. Yet, there is a second uncanny valley, the valley of imperfect perfection, which happens when AI transcends human abilities but not to the point of no contest [Boltuc 2017]. This second uncanny valley effect is the gist of the *singularity* fear among the fearful elites trying to slow down [Metzinger] or throttle developments in artificial intelligence. The slowdown of civilizational progress is the opposite of prudence in environmentalism [Bisk] and much more so in cognitive sciences. Not only Adam Smith, but also Karl Marx (in his crushing critique of Proudhon), have understood that attempts to strangle progress, however well-meaning attempts, have always been the opposite of humanism.

7 The Importance of findings

What is subject that is not an object good for?

Pure subjectivity is relevant for epistemic co-constitution of reality, as visible in [Fichte] and to some degree [Kant]. Without epistemic level the objects would never cross from potentiality to actual existence since, as Kant shows, objects in themselves have insufficient specifications.

The value of pure subject of first-person consciousness transpires most easily in second-person relationships with other consciousnesses, or epistemic monads [Leibniz, Buber, Levinas]. We can see this in the case of Church-Turing Lovers [Boltuc 2017]. Imagine that one can have one of the two significant others (companions and *lovers*), who are identical in any way but one, companion has first-person consciousness and the other does not (being a perfect humanoid machine or *a philosophical zombie*). Since we have existential/teleological reasons, of our lives making non-solipsistic sense, to care whether our significant have positive first-person experiences in the sense-making relationship, *ipso facto* we have reasons to care whether they have experiences at all. This is because we care about them – through love – but also because relationships with significant others are sense-making relations for ourselves.

If AGI is to dwell in the world of meaningful existence, it needs to have first-person consciousness. In principle this condition could be satisfied by cyborgization, but this wouldn't do since a machine with a borrow brains, say of a squirrel, would not be integrated with the borrowed consciousness in the relevant way. Even advanced brains would barely be the true consciousness of AGI, they would be just consciousness in the AGI and directing it would turn AGI into a gear, not a conscious entity.

If AGI is to have existence, not to dwell the shadows [Plato], it requires full epistemic subjectivity. This would amount to meeting the standards of the Engineering Thesis in Machine Consciousness [Boltuc 2009, 2012, 2007].

References *draft*

Abbott, R. B. (2017) Patenting the Output of Autonomously Inventive Machines; Landslide, Vol. 10, No. 1, September/October 2017, American Bar Association
https://www.americanbar.org/groups/intellectual_property_law/publications/landslide/2017-18/september-october/patenting-output-autonomously-inventive-machines/
Coulter, M. Patent agencies challenged to accept AI inventor *Financial Times* 08/31/2019

Block, N. (1995). ON A CONFUSION ABOUT A FUNCTION OF CONSCIOUSNESS. Behavioral and Brain Sciences 18 (2): 227-287

Boltuc 2009, 2012, 2007
Boltuc 2017,
Boltuc 2019 (Burgin)
Boltuc 2019b, 2019c (BICA)

Chalmers
Deutsch

Fichte,
Goertzel 2006,
Husserl
Kant

Kelley. T. "Developing a Psychologically Inspired Cognitive Architecture for Robotic Control: The Symbolic and Subsymbolic Robotic Intelligence Control System (SS-RICS" (2006, p. 219)

Hobbes
Libet
[Nagel 1979]
Thomas Nagel; View from Nowhere

[Russell 1921]

Shalom, A.

Thaler 2014
Thaler 2019 (APA)
[Turing 1950].